UCRL-JRNL-202059

Insights into the genome evolution of Yersinia pestis through whole genome comparison with Yersinia pseudotuberculosis

P.S.G. Chain, E. Carniel, F. Larimer, J. Lamerdin, W.M. Regala, A.M. Georgescu, L.M. Verguez, M. Land, B. Souza, L. Hauser, V.L. Motin, R. R. Brubaker, J. Fowler, J. Hinnebusch, M. Marceau, C. Medigue, M. Simonet, V. Chenal-Francisque, P.O. Stoutland, J.M. Elliott, A. Derbise, D. Dacheux, E. Garcia

January 28, 2004

Proceedings of the National Academy of Sciences (USA).

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Insights into the evolution of *Yersinia pestis* through whole genome comparison with *Yersinia pseudotuberculosis*

P. S. G. Chain[1], E. Carniel[2], F. W. Larimer[3], J. Lamerdin[1], P. O. Stoutland[1], W. M Regala[1], A. M. Georgescu[1], L. M. Vergez[1], M. L. Land[3], V. L. Motin[1], R. R. Brubaker[4], J. Fowler[4], J. Hinnebusch[5], M. Marceau[6], C. Medigue[7], M. Simonet[6], V. Chenal-Francisque[2], B. Souza[1], D. Dacheux[2], J. M. Elliott[1], A. Derbise[2], L. J. Hauser[3], and E. Garcia[1*]

[1]*Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA*
[2]*Yersinia Research Unit, Institut Pasteur, 75724 Paris Cedex 15, France*
[3]*Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[4]*Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA*
[5] *Rocky Mountain Laboratories, Hamilton, MT 59840, USA*
[6]*Inserm E0364 -Université de Lille 2-Institut Pasteur de Lille, F-59021 Lille, France*
[7]*Genoscope/CNRS-UMR 8030, 91006 Evry Cedex, France*

* Corresponding author
Emilio Garcia
BBRP, LLNL
7000 East Ave. L-452
Livermore, CA 94550
garcia12@llnl.gov
Ph. 925-422-8002
Fax 925-422-2099

23 text pages (including 2 Fig. legends), 1 Table, 2 Figures, 8 Supplemental Tables
Words in abstract: 195; Characters in paper: 46,850
Abbreviations: IS, insertion sequence
Data deposition: The sequences reported in this paper have been deposited in the GenBank database [accession nos BX936398 (chr.); BX936399 (pYV); BX936400 (pYptb32953)].

Yersinia pestis, the causative agent of plague, is a highly uniform clone that diverged recently from the enteric pathogen _Yersinia pseudotuberculosis._ Despite their close genetic relationship, they differ radically in their pathogenicity and transmission. Here we report the complete genomic sequence of _Y. pseudotuberculosis_ IP32953 and its use for detailed genome comparisons to available _Y. pestis_ sequences. Analyses of identified differences across a panel of _Yersinia_ isolates from around the world reveals 32 _Y. pestis_ chromosomal genes that, together with the two _Y. pestis_-specific plasmids, represent the only new genetic material in _Y. pestis_ acquired since the divergence from _Y. pseudotuberculosis._ In contrast, 149 new pseudogenes (doubling the previous estimate) and 317 genes absent from _Y. pestis_ were detected, indicating that as many as 13% of _Y. pseudotuberculosis_ genes no longer function in _Y. pestis_. Extensive IS-mediated genome rearrangements and reductive evolution through massive gene loss, resulting in elimination and modification of pre-existing gene expression pathways appear to be more important than acquisition of new genes in the evolution of _Y. pestis._ These results provide a sobering example of how a highly virulent epidemic clone can suddenly emerge from a less virulent, closely related progenitor.

Strong molecular evidence supports the fact that *Y. pseudotuberculosis*, responsible for yersiniosis in animals and humans, is the recent ancestor to *Y. pestis*, the etiologic agent of bubonic and pneumonic plague (1-3). However, while *Y. pseudotuberculosis* is a soil and water-borne enteropathogen, *Y. pestis* is much more dangerous and is of current interest due to its potential use in bioterrorism and as a biological weapon. Present day *Y. pestis* strains, though all similarly pathogenic, can be classified into three biovars (Antiqua, Medievalis, Orientalis) on the basis of their ability to utilize glycerol and to reduce nitrate. These phenotypic differences and molecular typing methods in conjunction with strain geographical origins have served to correlate these biovars with the three recorded plague pandemics.

Of special importance to the pathogenic process of both *Y. pseudotuberculosis* and *Y. pestis* is the shared requirement of a virulence plasmid pCD1 (pYV in enteropathogenic *Yersinia*) that encodes a type III secretion system (4), responsible for injecting into host cells a number of cytotoxins and effectors (Yops) that inhibit bacterial phagocytosis and processes of innate immunity (5, 6). Two additional plasmids unique to *Y. pestis* termed pPCP1 (9.6 kb) and pMT1 (102 kb) play roles in tissue invasion (7, 8) and capsule formation (9) as well as infection of the plague flea vector (10, 11), respectively. However, the presence of these plasmids by themselves cannot account for the remarkable increase in virulence observed in *Y. pestis* (12-15). Despite many extensive studies of the plasmid-encoded virulence determinants induced during the infectious process, and the recent availability of the genome sequences of a *Y. pestis* Orientalis strain, CO92 (16) and a Medievalis strain, KIM10+ (17), the mechanism(s) underlying the strikingly different clinical manifestations of *Y. pseudotuberculosis* and *Y. pestis* have remained elusive. Although a microarray-based comparison of these two Yersinia species has been reported recently (18), the detailed comparison between the completed genomes of *Y. pestis* and that of *Y. pseudotuberculosis* IP32953 (serotype I) presented here provides the first opportunity to examine all differences in genome structure and at the nucleotide level. These comparisons reveal many of the molecular details that were involved in the speciation and emergence of *Y. pestis* and may hold the key to the exceptional virulence of the plague bacillus.

Materials and Methods

Whole genome shotgun libraries were obtained by fragmenting genomic DNA using mechanical shearing, and cloning 2-4 kb fragments into pUC18. Double-ended plasmid sequencing reactions were carried out using BigDye Terminator chemistry (Perkin Elmer, Foster City, CA) resolved on ABI3700 Automated DNA Sequencers. The whole genome sequence of *Y. pseudotuberculosis* IP32953 was obtained from 85,000 end sequences (8.8 fold redundancy), and was assembled using PHRAP (P. Green, University of Washington). All gaps were closed by primer walking on gap-spanning clones or PCR products and a large insert scaffold was used to verify proper genome assembly. Gene modeling and genome annotation was performed as previously described (19). Genome comparisons between the *Yersinia* sequences were viewed using the Artermis Comparison Tool (ACT) (http://www.sanger.ac.uk/Software/ACT/).

The *Yersinia* strains studied came from the collection at the Institut Pasteur. Analysis of these strains was performed by screening PCR results and if necessary, sequencing the resulting products. Specifically, for the strain- or species-specific genes, primers were designed to amplify a ~500 bp region within the gene (or gene portion) that was found to be missing from the other strains or species. For the IS-interrupted genes, primers were designed to amplify a 300-600 bp region of the WT gene or a 1-2.5 kb fragment which includes the interrupting IS element. Due to homologous recombination between IS elements, the alternative and sometimes expected result was a negative one (e.g. no PCR product when the IS in question underwent recombination, or in the event of a deletion removing at least one of the priming sites). For the other pseudogenes, sequencing each PCR product was followed by multiple alignments of the sequences to identify wild-type versus mutant loci. In all cases, experiments yielding negative results were repeated under the same conditions and also using a lower annealing temperature in the event that the region in question had undergone divergence.

The *Y. pestis* and *Y. pseudotuberculosis* genomic DNAs that were used in panel-screens were isolated from the following strains of *Y. pestis*, Harbin (Former Soviet Union, biovar Antiqua), Japan (Japan, Antiqua), Margaret (Kenya, Antiqua), 343 (Belgium Congo, Antiqua), PKH-4 (Kurdistan, Medievalis), PKR292 (Kurdistan, Medievalis), PAR13 (Iran, Medievalis), 297RR (Vietnam, Orientalis), Exu184 (Brazil, Orientalis), Hambourg10 (Germany, Orientalis), 6/69 (Madagascar, Orientalis); and *Y. pseudotuberculosis*, IP33134 (Russia, serotype I), IP32790 (Italy, I), IP32950 (France, I), IP30215 (Denmark, II), IP32802 (Italy, III), IP32889 (Spain, III), IP31833 (England, IV), IP32952 (France, V). The controls used were *Y. pestis* CO92 (USA, biovar orientalis) and *Y. pseudotuberculosis* IP32953 (France, serotype I). Results for *Y. pestis* KIM10+ were predicted using the available genome sequence.

The sequence of the complete *Y. pseudotuberculosis* strain IP32953 genome is available under the following accession numbers: BX936398 (chromosome); BX936399 (pYV); BX936400 (pYptb32953). Supplementary information accompanies this paper and is accessible at www.pnas.org.

Results and Discussion

Genome organization of *Y. pseudotuberculosis* IP32953. The genome of strain IP32953, a fully-virulent clinical isolate from a human patient, consists of a single circular chromosome (4,744,671 bp), the pYV virulence plasmid (68,526 bp) and an atypical novel 27,702 bp cryptic plasmid designated pYptb32953. The general features of the IP32953 genome are listed in Table 1 and the chromosome is represented in Fig. 1. Comparisons of pYV to the previously sequenced pCD1 plasmids from *Y. pestis* KIM5 (20, 21) and *Y. pestis* CO92 (16) revealed an essentially conserved co-linear backbone, differing by the presence in pCD1 of an IS*100* element, an open reading frame (ORF) encoding a 68 amino acid hypothetical protein of unknown function, and an apparent internal in-frame 120 amino acid insertion in the middle of the *yopM* gene, consistent with the known heterogeneity found among YopM in yersiniae (22).

The novel pYptb32953 is likely a conjugative cryptic plasmid and bears similarity to the recently described cryptic plasmid of *Y. enterocolitica* strain 29930 (23). The similarity (~60-65 %) extends to the plasmid mobilization machinery involving TraE and MobB/C homologs, as well as to a cluster of genes encoding a type IV secretion system. The latter operon also displays homology to the conjugation genes of the IncX plasmid R6K (24) and to the *Brucella* spp. *virB* operon (25). Examination of the presence of this plasmid across a large number of isolates belonging to the three pathogenic *Yersinia* species indicated that its distribution is quite narrow (present in three out of 81 strains) (see Table 2 in supporting information). Thus this plasmid cannot account for any important virulence-associated characteristic of *Y. pseudotuberculosis*.

Few chromosomal features have been known to distinguish *Y. pestis* from *Y. pseudotuberculosis* strains (26, 27). However, comparisons between the chromosome of IP32953 and the *Y. pestis* chromosomes of CO92 and KIM10+ revealed several major differences (Table 1 and Fig. 1). The IP32953 chromosome encodes 3,974 predicted genes of which 2,976 (75%) have greater than or equal to 97% identity to their homologues in *Y. pestis*. Likewise, the synteny of the *Yersinia* genomes is readily discernable, and the breaks in colinearity have been mapped precisely.

Unique Chromosomal Regions in *Y. pseudotuberculosis*. Thirty-six IP32953-specific regions, ranging in size between 500 bp and 122 kb, are scattered throughout the chromosome and contain a total of 317 putative genes that are not found in either CO92 or KIM10+. A list of putative proteins encoded in these regions and their gene locations can be found in Table 3 (see supporting information). More than one half (188) of the genes in these unique regions are distributed in five clusters composed of phage-like products, the largest of which is a 122 kb composite phage region with high homology to bacteriophage P2. Another seven clusters encode 49 genes involved primarily in transposition and restriction modification and are likely horizontally-acquired. It appears that the remaining clusters encoding 80 genes have been deleted from the *Y. pestis* genome. The distribution of these genes (other than phage) into functional categories is shown in Fig. 2. Roughly a third of all the IP32953-specific genes in this group are hypothetical or conserved hypothetical genes, while others include genes that encode general metabolic functions that appear to have been lost in *Y. pestis*.

Since species-specific regions and other species-specific genomic features should be conserved across a broad section of strains, a panel of 19 geographically and phenotypically diverse strains of *Y. pestis* and *Y. pseudotuberculosis* was selected and screened for the presence or absence of these features. Of 85 IP32953-specific genes tested by PCR, 11 genes were found to be specific for the *Y. pseudotuberculosis* species (i.e. present in all *Y. pseudotuberculosis* and absent from all *Y. pestis* isolates – see Table

4 in supporting information). Only one of these genes, YPTB0537, lies within one of the 12 above-described regions of putative foreign origin. Four of these 11 genes encode hypothetical proteins while four others encode: a putative restriction modification system component (YPTB0537), and proteins involved in glucan biosynthesis (YPTB2493 and YPTB2494) and uracil transport (YPTB2793). The last three encode metabolism-related functions: aspartate aminotransferase, enolase-phosphatase E1, and 5-methylthioribose kinase, respectively. These differences in metabolic enzymes may reflect the differences in *Y. pestis* and *Y. pseudotuberculosis* host ranges. In addition, the *Y. pseudotuberculosis*-specific regions may account for important virulence factors uniquely required in *Y. pseudotuberculosis*, such as the *opg* operon (YPTB2493-YPTB2495) which is required for the synthesis of periplasmic branched glucans that serve in other organisms as an osmoprotectant (28) but that may not be needed in *Y. pestis*, an obligate parasite of eukaryotes, unlikely to experience wide fluctuations in environmental osmotic conditions.

Unique Chromosomal Regions in *Y. pestis*. We also identified 112 KIM10+ and CO92-specific (i.e. not found in IP32953) genes distributed in 21 clusters of 300 bp to 41.7 kb scattered throughout the genome (see Table 5 in supporting information). Roughly three categories of genes were identified in these 21 regions: 1) 39 genes (35% of the total) are hypothetical or conserved hypothetical; 2) 59 genes (53%) are phage or transposon-related and 3) 14 genes (12%) can be attributed a putative function. Among those with an ascribed function are membrane proteins, lipoproteins, a putative esterase, a DNA binding protein and a methyltransferase. Our studies indicate that a CO92 nine kb filamentous prophage region, previously believed to be Orientalis biovar specific (18, 27), is in fact also present in some members of the Antiqua biovar ((29) and see Table 6 in supporting information) and is absent from IP32953.

Of the 112 genes uniquely associated with the two *Y. pestis* genomes, 105 were tested for their presence or absence in our panel of 19 Yersinia strains (see Table 6). Only 32 genes, located in six clusters, were present in all *Y. pestis* and absent from all *Y. pseudotuberculosis* strains examined. Four of these clusters have been recently identified using microarray analysis (18). However, genome sequence comparison coupled with PCR has identified two additional regions not found by hybridization and has eliminated the five other regions previously determined as unique to *Y. pestis* (see Table 6) by that method.

Four of the *Y. pestis*-specific gene clusters encode predominantly putative proteins with little, if any, similarity to known or predicted proteins (with the exception of a methylase). Another cluster consists of bacteriophage-related genes (YPO2084-103, YPO2114 in CO92; y2227-y2211, y2201 in KIM10+); while the last cluster (YPO1668-71 in CO92; y1829-y1832) encodes putative membrane proteins, a translation initiation inhibitor, and conserved hypothetical proteins. Though there were no obvious virulence factors encoded in these regions, their role in pathogenicity deserves further study.

Inactivated Genes. Sixty-two pseudogenes are found in IP32953, 43 of which are also pseudogenes in one or both sequenced *Y. pestis* strains (see Table 7 in supporting information). The remaining 19 likely represent recent *Y. pseudotuberculosis*-acquired mutations that have arisen since their divergence. Of these, the functions most frequently affected included outer membrane, transport and exported proteins, perhaps reflecting the organism’s interaction with its environment. Two of the 62 were integrases with substantial similarity to one another: a P4-like integrase (YPTB0534) and the pathogenicity island HPI integrase (YPTB1602). Though the significance of these inactivated integrases in *Y. pseudotuberculosis* and the presence of intact counterparts in *Y. pestis* is not known, it is possible that they may be involved in the increased frequency of IS transposition in the latter.

Of the 149 originally reported CO92 pseudogenes (16), only 84 are pseudogenes in the KIM10+ strain and yet are intact genes in IP32953, suggesting a previous overestimation of the contribution of these pseudogenes to the phenotype of *Y. pestis*. Three-way gene by gene comparisons among the *Yersinia* strains enabled us to identify 149 additional putative pseudogenes in CO92 (see Table 8 in supporting information) of which 124 are also pseudogenes in the KIM10+ genome. Thus, a closer approximation to the factual number of potentially lost functions by this evolutionary mechanism in *Y. pestis* is 208 (84+124), so that as much as 5% of the gene complement may have been selectively inactivated in *Y. pestis*. A summary of this subset of inactivated genes and their distribution by COG functional classes is shown in Fig. 2.

Using the same panel of 19 strains, we also examined the distribution of 52 randomly-selected CO92 pseudogenes. Forty-six of them could be grouped into five discernable categories, the largest of which comprises 28 pseudogenes specific to *Y. pestis* (see Table 9 in supporting information). Members of this group are potentially the most interesting since they affect traits that are unique to *Y. pestis* strains and thus, may represent good targets for studying their novel pathogenic properties and for quick identification in clinical settings. Genes disrupted in this group range from conserved hypothetical, to genes of general metabolism such as *metB* (responsible for the observed methionine requirement of *Y. pestis*), to regulatory genes (e.g. putative two-component sensor kinase, etc.) and potential virulence-associated genes (invasin, toxin transporter, etc.).

A second group of seven pseudogenes was only found in members of the biovar Orientalis, and include the arginine-binding periplasmic protein 2 precursor (*argJ*); the N-terminal region of *E. coli* prepilin peptidase dependent protein (*ppdA*); the exonuclease encoded by *sbcC*; and the aerobic glycerol phosphate dehydrogenase (*glpD*), which is likely responsible for the glycerol-minus phenotype of the biovar Orientalis (30).

Six IS-interrupted pseudogenes comprise a third category, including *aroG*, *pbpC* (penicillin-binding protein 1C) and *setA*, a sugar efflux transporter. These are pseudogenes in all members of the Orientalis biovar as well as in one or both of the African Antiqua strains (from Kenya and Congo), and are intact genes in *Y. pseudotuberculosis*, the Medievalis lineage, and the non-African Antiqua strains. This finding, in addition to the previously alluded to filamentous phage distribution pattern supports the notion that the Orientalis and Medievalis lineages arose independently from Antiqua biovar.

A fourth category of two other pseudogenes, a putative surface protein (YPTB3178) and a pectin degradation protein (YPTB2355) are found in all *Y. pestis* strains and are also present in several *Y. pseudotuberculosis* strains. These may represent mutations acquired prior to the emergence of *Y. pestis*, as they are unlikely to have been independently acquired by each species (one is a partial deletion and the other is interrupted by an IS*285*).

The fifth and last category comprises a single IS*100*-interrupted acetylornithine aminotransferase, *argD*, a CO92-specific pseudogene, likely the result of a very recent insertion sequence mobility that supports the idea of a continuously fluid genome.

Metabolism. Since *Y. pseudotuberculosis* is a chemoheterotroph, a full complement of biosynthetic and intermediary metabolic pathways was expected and has been verified. As already indicated, several of the IP32953-specific regions encode general metabolic functions and thus, may account for some of the observed physiological differences between the two species. Noteworthy among this group are genes of purine and aspartate metabolism as well as of the methionine salvage pathway (31-33). Gene inactivations that may account for the *Y. pestis*-specific biochemical phenotypes include: a cysteine synthase (*cysM*) frameshift (the cysteine requirement of *Y. pestis*), a missense point mutation affecting amino acid 363 in the aspartate ammonia-lyase (*aspA*) of *Y. pestis* likely accounting for the stimulatory effect of CO_2 on growth (34), and a proline substitution present in amino acid 161 of glucose 6 P-dehydrogenase (*zwf*) that likely prevents utilization of hexose via the pentose-phosphate pathway (35). The significance of these last two types of mutations will require further functional analyses.

Pathogenicity. Genomic differences that may play a role in the unique pathogenic characteristics of these two species include alterations in lipid A biosynthesis exemplified by the absence in *Y. pestis* of lipid A acyltransferase gene *htrB* (YPTB2490), which adds an acyl group to lipid A, and may account for the differences in lipid A between the two species. Since lipid A acylation changes are known to alter endotoxic properties and interactions with the innate immune system, this difference could be of significance for pathogenesis.

Several ORFs with homology to hemolysins/hemagglutinins of different pathogens are present in the yersiniae. In IP32953, a cluster of nine ORFs (YPTB3450-YPTB3459) encode several hemolysin homologues in a region absent from *Y. pestis*. A hemolysin activator is a pseudogene in both IP32953 (YPTB3651) and KIM10+ (y0002) but is wild-type in CO92 (YPO3720). However, since this mutation occurs at a homopolymeric tract of C's (11 in IP32953 and KIM10+, and only 7 in CO92), it may simply represent a spontaneous reversion, similar to that shown to occur in *ureD* in which silencing and reactivation of urease in *Y. pestis* is determined by a spontaneous addition/excision of a single G residue in the *ureD* gene (36). Another hemolysin gene that is inactivated by partial deletion in IP32953 (YPTB2524) and all other *Y. pseudotuberculosis* strains is found intact in *Y. pestis* (see Table 6; gene YPO2486 in CO92 and y1701 in KIM10+). Although the role of hemolysins in *Yersinia* virulence remains unclear, their conserved nature and clear differences among the species suggest the need for further studies to investigate their possible function.

The insecticidal toxin homologs found in either complete or inactivated form in the *Y. pestis* genomes have been implicated in the adaptation of this organism to the flea life cycle (16, 18, 37). Thus, it has been suggested that the observed inactivation of *tcaB,* encoding an insecticidal toxin protein, is required for flea life cycle but this argument can now be refuted as this gene is complete and normal in several Medievalis and Antiqua *Y. pestis* strains (see Table 9). Similarly, the in-frame deletion of *tcaC* in *Y. pestis* cannot alone account for its ability to colonize the flea midgut, as this gene is even shorter in *Y. pseudotuberculosis*, neither can the same function be attributed to the viral enhancing protein previously described in CO92 (16), since it is also present in IP32953. Thus, the precise role of insecticidal toxin homologs in flea midgut colonization remains largely unresolved.

Two loci (*srfA* and *srfB*) encoding putative virulence factors, along with the gene for the Cu-Zn superoxide dismutase, *sodC* have in-frame insertion/deletions in KIM10+ and CO92, but are wild-type in IP32953. If these mutations affect protein function, they could play a role in species-specific virulence. Similarly, an IS*1541* neighboring *csrB,* a small non-coding RNA that antagonizes CsrA, an S-layer protein involved in adherence to cells, may modify the transcription and/or stability of this RNA and thus may have an effect on virulence in *Y. pestis.* Another region that could have a role in virulence in these organisms is a putative pathogenicity island (HPI 2). This region is wild-type in *Y. pseudotuberculosis* but defective in *Y. pestis* in which the siderophore synthesis protein (YPO0778 and YPO1012 in CO92; y3406 and y3410 in KIM10+) is inactivated by an IS*100* insertion.

Regulatory Genes. At least nine regulatory genes that are inactivated in *Y. pestis* could have effects on its phenotype, including virulence. A frameshift in *Y. pestis*-homologues of YPTB0553 affects a gene similar to *sorC*, a transcriptional regulator required for sorbose utilization, while a frameshift in the *Y. pestis* homologues of YPTB1259 may affect the regulation of the synthesis of polysaccharide colanic acid. This capsular polysaccharide has been implicated in blocking the specific binding between uropathogenic *E. coli* and inert substrates (38). These inactivations in *Y. pestis* may be consistent with the general loss of adhesins that are unnecessary for its life-style. The gene *flhD* may be one of many genes inactivated in *Y. pestis* responsible for altered motility in this organism. Its absence may have a positive impact on Yop expression (39) and a possible pleiotropic effect on virulence and metabolism as demonstrated in other enterobacteria (40-42). Also inactivated in *Y. pestis* is the *rhafR* homologue (YPO1728 in CO92; y2579 in KIM10+) which may lead to derepression of the rhaffinose utilization pathway in this organism.

A frameshift in the transcriptional regulator, *iclR* carried by *Y. pestis* leads to constitutive glyoxylate bypass in this organism explaining an already known phenomenon (43). Furthermore, since the glyoxylate bypass has been shown to be necessary for virulence in other bacterial pathogens and fungi (44, 45), constitutive expression may also enhance *Y. pestis* virulence.

UhpB (YPTB3846), a transcriptional activator of genes involved in the uptake and metabolism of hexose phosphates, is inactivated in many *Y. pestis* strains. Finally, the gene encoding sigma N modulating factor (YPTB3527) possesses a stop codon in position 36 in *Y. pestis*, which could lead to modified expression of sigma 54 dependent genes.

IS Elements, Genome Rearrangements and Evolution. Only 20 IS elements were found in the IP32953 chromosome, in stark contrast to the 117 in KIM10+ and 138 in CO92 (Table 1). Twelve of the 20 IS elements in IP32953 share integration locations with those in the two *Y. pestis* strains, suggesting that only 8 recent transposition events have occurred in IP32953, whereas an extraordinary expansion of each IS family took place in *Y. pestis* strains since their divergence. Examination of the shared IS locations within CO92 and KIM10+ suggests that their most recent common ancestor carried 109 IS elements and that since the divergence of this ancestral representative and the present day KIM10+ and CO92 strains, 8 and 28 new insertions occurred respectively. What remains unclear is whether the rate of transposition in *Yersinia* is periodically stimulated or if these events occurred in a punctuated fashion upon some as yet unknown induction.

Despite the dense distribution of IS elements in *Y. pestis* and their potential for generating homologous recombination-mediated deletions, there are surprisingly few (only five) IP32953-specific regions that can be the result of excision of intervening sequence via recombination at flanking direct IS elements in *Y. pestis*.

Deng *et al.* (17) first alluded to the important role played by repeat elements (namely IS elements) in explaining the unique genome arrangement displayed by the two sequenced *Y. pestis* strains. Analyses using the structural organization of IP32953 for comparison further support the role played by IS elements in genome evolution and confirms the ancestral character of *Y. pseudotuberculosis*, since IP32953 most often has no “equivalent” IS element when compared with *Y. pestis*. This implies that most rearrangements have occurred only recently in the *Y. pestis* lineage and that the genome structural organization of IP32953 more closely reflects that of the ancestral type.

In a manner analogous to that utilized in the KIM10+/CO92 comparisons (17), we can identify some 32 syntenic colinear blocks conserved between IP32953 and CO92 (Fig. 1) and 25 between IP32953 and KIM10+. The genome organization of the last common ancestral genome of the two *Y. pestis* strains, as well as the ancestral genome of both species, could be deduced by investigating the precise locations of these rearrangements. Thus, IP32953 has undergone at least one and likely no more than three intra-chromosomal recombinations since the split from the last common ancestor. A probable recombination in IP32953 that generated a large inversion between two IS*1661* is supported by the distinct shift in GC skew associated with this region (Fig. 1). Two other putative IP32953 rearrangements are exemplified by the mobile pathogenicity island HPI, which typically integrates at one of three *asn*-tRNAs in *Yersinia* spp. (46), and a recombination at a P4-like integrase (YPTB0534) that is common to all three sequenced strains. All other rearrangements appeared to have occurred in the *Y. pestis* lineage.

Allowing for the three possible rearrangements proposed during the evolution of IP32953, the progenitor of both CO92 and KIM10+, must have undergone at least 11 recombination/rearrangement events (undoubtedly influenced by the 97 additional IS elements gained since diverging from *Y. pseudotuberculosis*). KIM10+ and CO92 have since undergone an additional 10 and 18 rearrangements respectively, commensurate to their respective increased levels of IS transposition. It is thus quite likely that the insertion elements themselves and/or the subsequent rearrangements they have generated have played an important role in the emergence of *Y. pestis* from its *Y. pseudotuberculosis* ancestor.

Implications in virulence and pathogen evolution

The genome sequence of *Y. pseudotuberculosis* IP32953 and its comparison to *Y. pestis* reveal aspects of the evolutionary processes that evidently transformed a common enteropathogenic ancestor, and later gave rise to two present-day pathogens of vastly distinct clinical manifestations. Molecular events that likely operated during the evolution of alternatively free-living *Y. pseudotuberculosis* (capable of causing localized chronic disease) contrast with those involved in the evolution of *Y. pestis* (capable of causing vector-dependent acute disease). The extensive chromosomal rearrangements that occurred during the emergence of *Y. pestis* undoubtedly are indicative of the mechanisms that drove the evolution of this pathogen. IS element expansion and its corollary, the increased fluidity of the genome, together with massive gene inactivation almost surely have played a role in this process. A direct comparison between the work presented in this study and the calculated evolutionary distances between these two *Yersinia* species presented by Achtman *et al*. (1) is difficult to make without a reliable molecular clock to measure the rates of genome rearrangement and IS transposition and gene inactivation. Since the mechanism(s) that account for IS element expansion and increased gene inactivation in *Y. pestis* is unknown, we can only surmise that (after the lateral transfer of pPCP1 and pMT1) these processes were driven by selection for lethality as well as evolutionary pressures that further enabled colonization of the flea. In this scenario, the dramatic change in lifestyle undergone recently by *Y. pestis*, reflecting its continuous association with the host and dependency on the flea vector for survival, was sufficient to provide the selective pressure that resulted in wholesale inactivation of as much as 13% of its genome. This may represent an intermediate stage in genome compaction, a process that has been proposed in the evolution of other pathogens closely associated with their hosts such as *Salmonella typhi* (47) and *Mycobacterium lepae* (48). Finally, the significance of horizontal gene transfer into the chromosome of *Y. pestis* is uncertain. It may be hypothesized that the acquisition of at least some of the six chromosomal regions uniquely conserved in *Y. pestis* strains, in conjunction with the high degree of gene inactivation has been responsible for the increased pathogenicity of this species. Whole genome comparisons of pathogen near-neighbors of distinct characteristics, such as those described in this study, lay the foundation for future mutational, functional and animal studies that will ultimately help elucidate the mechanisms underlying the emergence of new pathogens.

We thank S. Falkow, D. Monack, P. Agron, L. Radnedge, M. Chu and C. Kim for helpful discussions and review of the manuscript. This work was performed under the auspices of the U.S. DOE by Lawrence Livermore National Laboratory under Contract W-7405-ENG-48 and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725, and by contract 99 01 110 00 470 94 50 from the French Délégation Générale à l'Armement (DGA).

1. Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. & Carniel, E. (1999) *Proc Natl Acad Sci U S A* **96,** 14043-8.
2. Brenner, D. J., Steigerwalt, A. G., Falcao, D. P., Weaver, R. E. & Fanning, G. R. (1976) *Int. J. Syst. Bacteriol* **26,** 180-194.
3. Moore, R. L. & Brubaker, R. R. (1975) *Int. J. Syst. Bact.* **25,** 336-339.
4. Cornelis, G. R. & Van Gijsegem, F. (2000) *Annu Rev Microbiol* **54,** 735-74.
5. Brubaker, R. R. (2003) *Infect Immun* **71,** 3673-81.
6. Cornelis, G. R. (2002) *Nat Rev Mol Cell Biol* **3,** 742-52.
7. Brubaker, R. R., Beesley, E. D. & Surgalla, M. J. (1965) *Science* **149,** 422-424.
8. Lahteenmaki, K., Virkola, R., Saren, A., Emody, L. & Korhonen, T. K. (1998) *Infect Immun* **66,** 5755-62.
9. Kutyrev, V. V., Popov Iu, A. & Protsenko, O. A. (1986) *Mol Gen Mikrobiol Virusol***,** 3-11.
10. Hinnebusch, B. J., Rudolph, A. E., Cherepanov, P., Dixon, J. E., Schwan, T. G. & Forsberg, A. (2002) *Science* **296,** 733-5.
11. Hinnebusch, B. J. (2003) *Adv Exp Med Biol* **529,** 55-62.
12. Filippov, A. A., Solodovnikov, N. S., Kookleva, L. M. & Protsenko, O. A. (1990) *FEMS Microbiol Lett* **55,** 45-8.
13. Friedlander, A. M., Welkos, S. L., Worsham, P. L., Andrews, G. P., Heath, D. G., Anderson, G. W., Jr., Pitt, M. L., Estep, J. & Davis, K. (1995) *Clin Infect Dis* **21 Suppl 2,** S178-81.
14. Kutyrev, V., Mehigh, R. J., Motin, V. L., Pokrovskaya, M. S., Smirnov, G. B. & Brubaker, R. R. (1999) *Infect Immun* **67,** 1359-67.
15. Welkos, S. L., Andrews, G. P., Lindler, L. E., Snellings, N. J. & Strachan, S. D. (2004) *Plasmid* **51,** 1-11.
16. Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B., Sebaihia, M., James, K. D., Churcher, C., Mungall, K. L., *et al.* (2001) *Nature* **413,** 523-7.
17. Deng, W., Burland, V., Plunkett, G., 3rd, Boutin, A., Mayhew, G. F., Liss, P., Perna, N. T., Rose, D. J., Mau, B., Zhou, S., *et al.* (2002) *J Bacteriol* **184,** 4601-11.
18. Hinchliffe, S. J., Isherwood, K. E., Stabler, R. A., Prentice, M. B., Rakin, A., Nichols, R. A., Oyston, P. C., Hinds, J., Titball, R. W. & Wren, B. W. (2003) *Genome Res* **13,** 2018-29.
19. Chain, P., Lamerdin, J., Larimer, F., Regala, W., Lao, V., Land, M., Hauser, L., Hooper, A., Klotz, M., Norton, J., *et al.* (2003) *J Bacteriol* **185,** 2759-73.
20. Hu, P., Elliott, J., McCready, P., Skowronski, E., Garnes, J., Kobayashi, A., Brubaker, R. R. & Garcia, E. (1998) *J Bacteriol* **180,** 5192-202.
21. Perry, R. D., Straley, S. C., Fetherston, J. D., Rose, D. J., Gregor, J. & Blattner, F. R. (1998) *Infect Immun* **66,** 4611-23.
22. Boland, A., Havaux, S. & Cornelis, G. R. (1998) *Microb Pathog* **25,** 343-8.
23. Strauch, E., Goelz, G., Knabner, D., Konietzny, A., Lanka, E. & Appel, B. (2003) *Microbiology* **149,** 2829-45.
24. Nunez, B., Avila, P. & de la Cruz, F. (1997) *Mol Microbiol* **24,** 1157-68.

25. Boschiroli, M. L., Ouahrani-Bettache, S., Foulongne, V., Michaux-Charachon, S., Bourg, G., Allardet-Servent, A., Cazevieille, C., Lavigne, J. P., Liautard, J. P., Ramuz, M. & O'Callaghan, D. (2002) *Vet Microbiol* **90,** 341-8.

26. Brubaker, R. R. (2000) in *The prokaryotes, an evolving electronic resource for the microbiological community, vol. 2000.*, eds. Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackelbrandt, E. (Springer Verlag, New York.).

27. Radnedge, L., Agron, P. G., Worsham, P. L. & Andersen, G. L. (2002) *Microbiology* **148,** 1687-98.

28. Kennedy, E. P. (1996) in *Escherichia coli & Salmonella*, ed. Neidhardt, F. C. (ASM Press, Washington DC), Vol. I, pp. 1064-1071.

29. Gonzalez, M. D., Lichtensteiger, C. A., Caughlan, R. & Vimr, E. R. (2002)*J Bacteriol* **184,** 6050-5.

30. Motin, V. L., Georgescu, A. M., Elliott, J. M., Hu, P., Worsham, P. L., Ott, L. L., Slezak, T. R., Sokhansanj, B. A., Regala, W. M., Brubaker, R. R. & Garcia, E. (2002) *J Bacteriol* **184,** 1019-27.

31. Mortlock, R. P. & Brubaker, R. R. (1962) *J Bacteriol* **84,** 1122-1123.

32. Brubaker, R. R. (1970) *Infect Immun* **1,** 446-454.

33. Dreyfus, L. A. & Brubaker, R. R. (1978) *J Bacteriol* **136,** 757-64.

34. Baugh, C. L., Lanham, J. W. & Surgalla, M. J. (1964) *J Bacteriol* **88,** 553-8.

35. Mortlock, R. P. (1962) *J Bacteriol* **84,** 53-9.

36. Sebbane, F., Devalckenaere, A., Foulon, J., Carniel, E. & Simonet, M. (2001) *Infect Immun* **69,** 170-6.

37. Wren, B. W. (2003) *Nat Rev Microbiol* **1,** 55-64.

38. Hanna, A., Berg, M., Stout, V. & Razatos, A. (2003) *Appl Environ Microbiol* **69,** 4474-81.

39. Bleves, S., Marenne, M. N., Detry, G. & Cornelis, G. R. (2002) *J Bacteriol* **184,** 3214-23.

40. Pruss, B. M., Campbell, J. W., Van Dyk, T. K., Zhu, C., Kogan, Y. & Matsumura, P. (2003) *J Bacteriol* **185,** 534-43.

41. Pruss, B. M., Liu, X., Hendrickson, W. & Matsumura, P. (2001) *FEMS Microbiol Lett* **197,** 91-7.

42. Pruss, B. M. & Matsumura, P. (1996) *J Bacteriol* **178,** 668-74.

43. Hillier, S. & Charnetzky, W. T. (1981) *J Bacteriol* **145,** 452-8.

44. Lorenz, M. C. & Fink, G. R. (2002) *Eukaryot Cell* **1,** 657-62.

45. Lorenz, M. C. & Fink, G. R. (2001) *Nature* **412,** 83-6.

46. Buchrieser, C., Brosch, R., Bach, S., Guiyoule, A. & Carniel, E. (1998) *Mol Microbiol* **30,** 965-78.

47. Parkhill, J., Dougan, G., James, K. D., Thomson, N. R., Pickard, D., Wain, J., Churcher, C., Mungall, K. L., Bentley, S. D., Holden, M. T., *et al*. (2001) *Nature* **413,** 848-52.

48. Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., Harris, D., *et al*. (2001) *Nature* **409,** 1007-11.

Fig. 1. Circular genome map of IP32953 and comparison with *Y. pestis* CO92.

Panel A, Genome of IP32953; Panel B, Genome of CO92. For both panels, circle 1 (from center outward), G+C content; circles 2 and 3, all genes coded by function (forward and reverse strand); circle 4, GC skew (G-C/G+C); circles 5 and 6, genome divided into locally colinear blocks (when compared with one another, IP32953 and CO92), each colinear block is distinguished by a unique color (black segments within colored blocks represent regions specific to that genome in this comparison) and the orientation of each block in indicated by strand (circle 5, +ve strand; circle 6, -ve strand); circle 7, locations of IS elements (blue IS*100*, red IS*285*, green IS*1661* , magenta IS*1541*). In panel A, the gray highlighted region near 12 o'clock indicates the proposed IP32953 inversion (see text) while the reminder of the genome denotes the stable "ancestral" arrangement that has prevailed through the present. Panel B illustrates the complexity of the molecular events that gave rise to the inversions or translocations in the *Y. pestis* genome first proposed solely on the basis of the dramatic shifts in G/C skew (gray highlights I, II and III) (16)but now is extended through whole genome comparison. For example, gray highlight II is composed of three distinct blocks, two derived from distinct places within the same replichore (origin to terminus half) while the third one originated from the other replichore (light blue block).

Fig. 2. Functional classification of genes missing or inactivated in *Y. pestis*. Distribution of *Y. pestis*-specific lost functions by gene region deletion (light blue) or by gene inactivation (i.e. pseudogene, dark purple) in COG functional groups: C, Energy production; D, Cell division, chromosome partitioning; E, Amino acid metabolism; F, Nucleotide metabolism; G, Carbohydrate metabolism; H, Coenzyme metabolism; I, Lipid metabolism; J, Translation; K, Transcription; L, DNA replication, repair; M, Cell envelope biogenesis; N, Cell motility, secretion; O, Posttranslational modification; P, Inorganic ion metabolism; R, General function prediction only; S, Function unknown; T, Signal transduction; conserved, conserved hypothetical genes with no significant COG hits; unique, hypothetical genes with no significant COG hit.